# Impact of Video Editing Based on Participants' Gaze in Multiparty Conversation

**Yoshinao Takemae[†], Kazuhiro Otsuka[†], and Naoki Mukawa[*]**

NTT Communication Science Laboratories, NTT Corporation[†]
3-1 Morinosato Wakamiya, Atsugi-shi,
Kanagawa, 243-0198 Japan
{takemae,otsuka}@eye.brl.ntt.co.jp

Tokyo Denki University[*]
2-1200 Muzai Gakuendai, Inzai-shi,
Chiba, 270-1382 Japan
mukawa@sie.dendai.ac.jp

## ABSTRACT

This paper presents a video cut editing rule based on participants' gaze for establishing video editing rules that can accurately and clearly convey the flow of conversation in multiparty conversations to viewers. Demand is growing to be able to effectively archive meetings and teleconferences to facilitate human communication. Conventional systems use fixed-viewpoint cameras and simple camera selection based on participants' utterances etc. However, these systems fail to convey a sufficient amount of nonverbal information about the participants and the flow of conversation. On the basis of participants' gaze behavior in multiparty conversation, we propose a new video cut editing rule; the rule uses majority decision with regard to participants' gaze direction. We then present experiments that compare the proposed method to conventional visual representations. We conclude that the proposed method can more successfully convey *1) who is talking to whom* and *2) hearers' response to speakers*, which are extremely crucial pieces of information that allow viewers to understand the flow of conversation.

**Categories & Subject Descriptors:** H.5.3 [Information interfaces and presentation (e.g., HCI)]: Group and Organization Interfaces – Computer-supported cooperative work

**General Terms:** Human Factors; Theory

**Keywords:** Video editing techniques; archiving meetings; teleconferencing; gaze; multiparty conversation

## INTRODUCTION

Meetings are one of the most important activities in many workgroups. Often, due to scheduling conflicts or travel constraints, some cannot attend their scheduled meetings. We can overcome these problems by archiving the meetings and teleconferences. The need for systems that can effectively archive such sessions is increasing. This research also addresses an important topic in the field of Computer Supported Cooperative Work (CSCW).

Our purpose is to establish video editing rules that can accurately and clearly convey the flow of conversation in multiparty conversations to viewers afterward. To this end, we focus on two fundamental components: *1) conversation direction* which shows who is talking to whom, and *2) hearer's response* to speakers, including silence. These localized components are extremely crucial pre-conditions in conveying the flow of conversation. The reason for this is that conversation is constructed from a series of pairs, each of which is a speaker's utterance and a hearer's response to the utterance.

In previous work we proposed a video cut editing rule based on the convergence of participants' gaze in multiparty conversations [1]. This novel approach exploits participants' gaze behavior to select the most effective shots of participants. Experiments on three-participant situations indicated the possibility of clearly conveying the conversational direction. However, the quantitative analysis of the relations between participants' gaze behavior and conversation direction etc. was insufficient. Moreover, evaluations using more than three participants and visual representations other than camera selection were not conducted.

In this paper, we propose a new video cut editing rule that supports more than three participants; the use of majority decision of participants' gaze direction is a major innovation in this paper. We analyze participants' gaze behavior in 3- to 5-participant conversations. We then conduct experiments to verify the effectiveness of the proposed method by comparing visual representations such as multiple view shots. The following sections summarize related work and introduce the proposed method. We present an analysis of participants' gaze pattern and the details of our experiments, and discuss the results.

## RELATED WORK

While this study focuses on archiving meetings and watching them afterward, a considerable overlap exists between this domain and teleconferencing. Most conventional systems use a fixed-viewpoint camera. In large multiparty situations, participant face size is small. Hence these systems cannot sufficiently convey nonverbal information such as changes in facial expressions and gaze. These visual cues greatly contribute to the viewers' understanding of participants' intention and emotion. Other conventional systems use visual representations that arrange multiple participants' shots captured by multiple cameras on one display. However these systems cause cognitive loads on the viewer who must select video windows, and so they hinder understanding of the conversation.

The solution to this problem is automatic camera selection in which multiple video streams of multiple participants are appropriately ordered before being distributed. Cluster et al. developed a system called ”Distributed Meetings” [2]. The system employs camera selection based on participants' utterances in addition to a panorama view shot. However, this approach cannot adequately convey whom a speaker is talking to and hearers' responses such as rigid face with silence etc, since only the speaker is shown. Inoue et al. proposed

*CHI 2004*, April 24–29, 2004, Vienna, Austria.
ACM 1-58113-703-6/04/0004.

a camera selection scheme based on a probability model obtained by analyzing the duration and the transition of shots in debate programs on TV [3]. This method provides viewers with video sequences that show speakers' shots only or other participants' shots. However this approach fails to convey the flow of actual conversations because it uses a probability model, which has no relation to the actual conversations.

For conveying the flow of conversation in TV programs and films, a number of cutting techniques are often used. Cutting technique is equivalent to camera selection. Approaches such as ”A Theory of Montage” [4] and ”Grammar of Film Language” [5] allow discontinuous shots to be formed into a montage that hopefully expresses the flow of conversation. By controlling the viewers' attention, they allow viewers to actively interpret and discern the relations between shots. For this reason cutting techniques such as ”L Cutting” and ”Shot/Reverse Shot” are used to handle conversations. Such cutting techniques reflect the experience of professional video directors and editors, and it is difficult for computers to completely reproduce their acquired knowledge.

In the area of cognitive psychology, the psychological impact of video editing techniques on viewers has been investigated. Reeves and Nass reported that cutting techniques trigger visual orientation and focus the attention of the viewers [6]. Most studies, however, did not focus on conversational scenes and failed to provide video editing guidelines for conveying the flow of conversation.

## PROPOSED METHOD

We use the convergence of participants' gaze direction to establish a video cut editing rule that can be applied to situations with more than three participants. Our previous rule selects a close shot of the participant that all participants, $(N-1)$, are looking at [1]. $N$ denotes the number of participants. This is based on the following assumptions: 1) A person gazes at another when that person is of interest: participants try to acquire visual cues such as facial expression and the gaze direction of the other participants, and interpret others' intention and emotion. This gaze behavior is called ”the monitoring function of gaze” [7]. 2) A person who receives the gaze of more participants has more important information with regard to the conversation.

However, in large multiparty situations, this rule fails to provide effective camera selection. The reason is that as the number participants increases, the gaze pattern becomes more dispersed and fleeting. To solve this problem, our new video cut editing rule, which offers the effectiveness of gaze convergence, is based on majority decision: the close shot is of the participant that most participants are gazing at. This principle can be expected to support meetings with more than three participants.

## ANALYSIS

We analyzed participant gaze patterns in 3- to 5-person conversations. As shown in Figure 1, conversation consists of a series of pairs, each of which is a unit interval of a speaker utterance *(U1)* followed by that of a hearer response *(U2)*. Hearer responses fall into two patterns: response with utterance, and that with silence involving changes of gaze and facial expression etc. To verify the effectiveness of the proposed method, we analyzed the time transitions of participant gaze direction, focusing on *(U1)* and *(U2)*. The following items were analyzed:

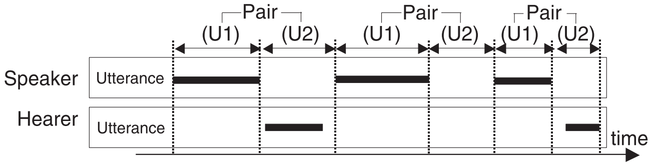

**Figure 1: A series of (U1) and (U2) pairs.**
The bold line shows the interval of the person making the utterance. (U1) and (U2) show the unit interval of speaker utterance and that of hearer response, respectively.

*(P1)* Within each unit interval, we calculated, as a percentage, how long participant gaze was focused on the speaker (or the hearer). We then calculated the average percentage for each.

*(P2)* We counted the number of unit intervals wherein the focus of participants' gaze switched between the hearer and the speaker or vise versa at least twice, and then calculated the proportion relative to all unit intervals.

### Collecting Conversational Data

We focused on 3- to 5-participant debates. Three groups with three participants, one with four participants, and one with five participants participated in the debates. Close shots of each participant and a whole view shot were recorded. Each group debated about topics such as whether we should accept euthanasia. Pin microphones recorded the utterances. All captured debates took about 120 minutes. Participant gaze direction was extracted manually with the temporal resolution of 33 ms from captured videos. Since we can predict the development of an automatic eye gaze tracker [8], all process can be automated. Utterance intervals of participants were automatically extracted based on power information of recorded voice. The speaker and hearer pairs were identified using voice power and speaker's gaze direction. 1248 *(U1)* and *(U2)* pairs were analyzed.

### Gaze Pattern in Multiparty Conversation

*1) Results of (U1) analysis. (P1)* gaze was directed at the speaker (70$\%$) and to the hearer (23$\%$). *(P2)* was 79$\%$.

*2) Results of (U2) analysis. (P1)* gaze was directed at the speaker (30$\%$) and to the hearer (51$\%$). *(P2)* was 45$\%$.

The results indicated that participant gaze focused on speakers as well as hearer in both utterance and response intervals. Therefore, the focus of participant gaze is a good way to identify conversation direction and hearer's response. These components can be extracted by analyzing the participant gaze patterns to some extent; there is no need for language analysis. Since the proposed method selects the participant who receives the gaze of most participants, it is expected to produce a video produced that can more accurately and clearly convey conversation direction and hearer's response to viewers.

Figure 2 shows gaze pattern in one part of a conversation and an example of videos sequences based on the proposed method (G). A pair of speaker utterance and wordless hearer response is given. Participant gaze alternated between the speaker and the hearer as shown in the gaze pattern of Figure 2. Consequently (G) caused the alternate camera selection of the speaker and the hearer.

## EXPERIMENTS

We conducted the experiments to verify the effectiveness of the proposed method. Subjects, who did not participate in the debates, viewed the resulting videos, and evaluated them. The accuracy and clarity of the videos in conveying the con-

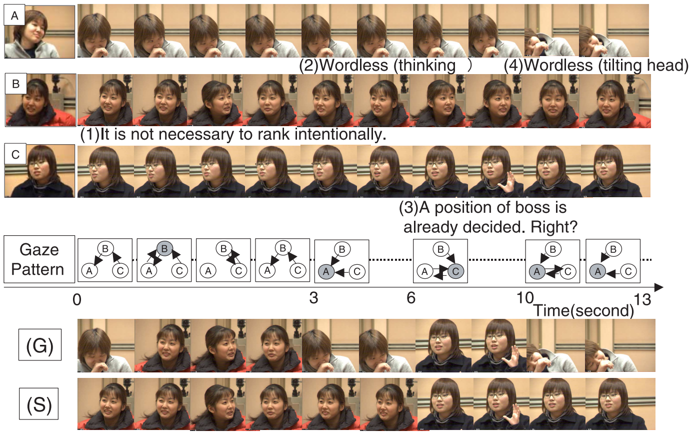
A
(2)Wordless (thinking )
(4)Wordless (tilting head)
B
(1)It is not necessary to rank intentionally.
C
(3)A position of boss is already decided. Right?
Gaze Pattern
0
3
6
10
13
Time(second)
(G)
(S)

**Figure 2: Gaze pattern in one part of a debate and examples of videos produced.**

In the gaze pattern, arrows indicate each participant's gaze direction. The person being gazed at by the two other people is shown in gray. (G) and (S) show the video sequence produced by the proposed method and the speaker shot, respectively.

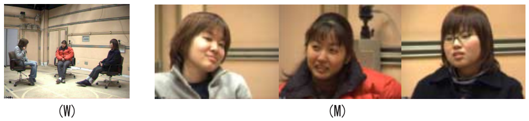

**Figure 3: Examples of whole view shot (W) (left) and multiple view shot (M) (right).**

versational direction and the hearer responses were evaluated. To this end, we focused on the fundamental unit intervals *(U1)* and *(U2)*. Two experiments were conducted:

*1) Experiment 1.* The conversational direction was evaluated using *(U1)*. The occasions in which the speaker was talking to one specific person were used.

*2) Experiment 2.* The conversation direction and hearer response were evaluated using pairs of *(U1)* and *(U2)*. Hearer responses were classified into two patterns: response with utterance, and that with silence. Each response was evaluated.

### Visual Representations Compared

The proposed method was compared to the following three visual representations, All of which are currently used for archiving meetings.

*1) Whole view shot (W).* All participants are captured in one shot as shown in Figure 3 (W). In large multiparty conversations, this cannot adequately convey the changes in facial expressions and gaze because participant face size is too small.

*2) Multiple view shot (M).* This places close shots in one row in order to express the spatial relations between participants (see Figure 3 (M)). This does not completely preserve the geometric arrangement of participants and makes it difficult for viewers to recognize whom the speaker is gazing at.

*3) Speaker shot (S).* The moment a participant starts an utterance, a close shot of the speaker is shown. Figure 2 (S) shows a video sequence produced by this method. This has the effect of clearly conveying who is the speaker.

## EXPERIMENT 1

### Method

*Subjects.* The paid subjects, who did not participate in the debates, were 57 people. Subjects were japanese in twenties. Subjects were divided into four groups. The first group, the second group, the third group, the fourth group viewed each of visual representations produced using approaches (G), (W), (M), and (S). The number of subjects in the groups was 15, 13, 14, and 15, respectively. These subjects also participated in experiment 2.

| Question No. | Questions |
|---|---|
| Q1-1-1 | Who do you think the speaker was ? |
| Q1-1-2 | Did you clearly see who the speaker was ? |
| Q1-2-1 | Whom do you think the speaker was talking to ? |
| Q1-2-2 | Did you clearly see whom the speaker was talking to ? |

**Table 1: Questionnaire for experiment 1.**

| Question No. | Questions |
|---|---|
| Q2-1-1 | Who do you think the first speaker was ? |
| Q2-1-2 | Did you clearly see who the first speaker was ? |
| Q2-2-1 | Whom do you think the first speaker was talking to at the end of his/her utterance ? |
| Q2-2-2 | Did you clearly see whom the first speaker was talking to ? |
| Q2-2-3 | What kind of response do you think the hearer showed to the speaker's question ? |
| Q2-2-4 | Did you clearly see how the hearer reacted to the speaker's question ? |

**Table 2: Questionnaire for experiment 2.**

*Materials.* The scenes, including continuous utterances of more than 4 seconds, were extracted from captured conversational data. Each scene was edited using the four different visual representations. The number of scenes in the 3-person, 4-person, and 5-person conversations were 6, 6, and 3, respectively.

*Questionnaire.* To evaluate conversation direction in *(U1)*, a questionnaire was used (See Table 1). Q1-1-1 and Q1-1-2 determine, respectively, the accuracy and clarity of recognizing the speaker. They also determine, respectively, the accuracy and clarity of recognizing whom the speaker is talking to: the hearer. In items of accuracy, the subjects were instructed to select the applicable person. In items of clarity, subjects selected one statement from a 7-point scale: -3 (strongly disagree) to 3 (strongly agree).

*Correct Answers.* The correct answer of Q1-1-1 was defined as the person, who was extracted from power information of voice. That of Q1-2-1 was defined as follows. Two evaluators, who did not participate in the debates, viewed videos of the entire debates any number of times, and subjectively determined the person (hearer) the speaker was talking to in each interval. Only scenes in which their evaluations accorded perfectly were employed. Experiment 2 was also conducted in the same way.

*Procedures.* Before starting the experiments, the subjects were instructed to memorize the participants' face and the spatial relations between participants; they also viewed videos of each participant speaking individually. After that, subjects in each group viewed a video based on each representation whose duration was one unit interval, and answered the questionnaire. This was one trial and multiple trials were conducted using different videos.

### Results and Discussion

The effects of participant number in the debates and visual representation style were analyzed using the data from the questionnaires. We used two-factor ANOVA with visual representation type and participant number as independent variables. If a significant difference was found, Tukey's multiple comparison was applied.

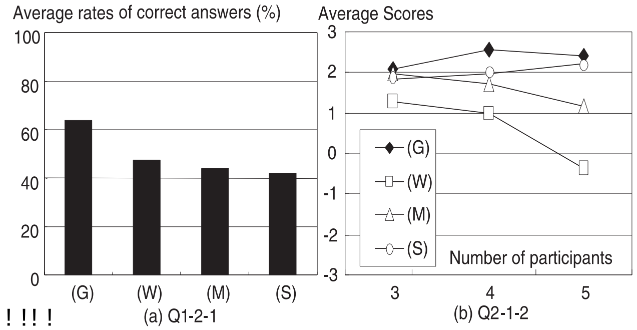

**Figure 4: Survey results.**

! ! (a) shows average rates of correct answers in Q1-2- 1 for experiment 1(left). (b) shows average scores in Q2-1-2 for experiment 2(right). The horizontal axis plots participant number (right). (G), (W), (M), and (S) represent the proposed method, whole view shot, multiple shot, and speaker's shot, respectively.

*Identification of hearer.* Figure 4 (a) shows average rates (%) of correct answers for each representation as determined from Q1-2-1. No significant impact was found with regard to participant number. The average rate of (G) was 65%, and less than 50% for the others. (G) was evaluated more highly than others ($p < .01$). Q1-2-2 yielded the same results ($p < .01$). In other words, the proposed method is more effective than the other representations for accurately and clearly conveying the hearer. This is because the shots of the speaker and hearer were alternately shown, based on participant gaze. In contrast, it was difficult for the subjects to recognize the speaker's gaze direction in (W) because face size was too small while (M) provided insufficient geometric coordination between participants' shots. (S) failed to show any shot of the hearer, because only the speaker was shown.

In addition, (G) yielded shots in which the speaker and the hearer were looking in opposite directions. This is called ”Matching the look” within ”Grammar of the Film Language” [5]. It is entirely intuitive that two people holding a conversation will gaze at each other and hence in opposite directions. This makes the shots produced by the proposed method more natural and effective.

## EXPERIMENT 2

### Method

*Materials.* The scenes, in which a speaker was demanding an answer from another, were extracted manually using language information. 15 scenes with responses, including silence, were used.

*Questionnaires.* We designed a questionnaire to evaluate the conversation direction and hearer response in *(U1)* and *(U2)* (see Table 2). Q2-1-1 and Q2-1-2 determine, respectively, the accuracy and clarity with which the speaker is recognized. Q2-2-1 and Q2-2-2 also determine, respectively, the accuracy and the clarity with which the person being addressed by the speaker at the end of his/her utterance was identified. Q2-2-3 and Q2-2-4 determine, respectively, the accuracy and the clarity with which the hearer's response to the speaker was identified. Q2-2-3 took the form of a multiple choice: agreement or disagreement.

### Results and Discussion

*Identification of speaker.* Figure 4 (b) shows average scores (from Q2-1-2) for each visual representation. Those of (G) and (S) held 2-point scores, even as participant number was increased. The scores of (W) and (S), on the other hand, fell ($p < .01$). In 5-person debates, (G) and (S) were significantly different from the others ($p < .01$). This indicates that (G) and (S), which are based on camera selection, are more effective than the others, if participant number is high. We expect that these trends will only be strengthened if participant number exceeds five.

*Identification of hearer and his/her response.* As noted above, hearer response has two patterns. As for responses with utterance, no significant difference was found among the representations. This is because only the utterance was analyzed. As an example of a response with silence, one participant (hearer) reacted negatively to the speaker, and the other participants gazed at the hearer to get visual cues such as facial expressions. For the responses with silence, (G) was evaluated more highly than others. Exline suggested that participants tend to gaze strongly at others in competitive situations [9]. The characteristic of this gaze behavior resulted in effective camera selection.

## CONCLUSIONS AND FUTURE WORK

This paper has proposed a video cut editing rule based on the majority decision of participants' gaze for conveying conversation direction and hearer's response to viewers. We analyzed the gaze behavior of participants in multiparty conversations and used the results to create the proposed method. We conducted experiments to compare the effectiveness of the proposed method against existing visual representation schemes for teleconferences. We conclude that videos produced by the proposed method can more accurately and clearly convey the conversation direction and hearer's response. This work offers a realistic framework for video editing based on the cue of participant gaze. In future work, we will focus on the development of a way to evaluate the flow of conversation and an automatic eye gaze tracker.

## REFERENCES

1. Y. Takemae, K. Otsuka, and N. Mukawa, Video Cut Editing Rule Based on Participants' Gaze in Multiparty Conversation, *Proc. of ACM Multimedia '03*, pp.303-306, 2003.
2. R. Cutler, et al., Distributed Meetings: A Meeting Capture and Broadcasting System, *Proc. of ACM Multimedia '02*, pp.503-512, 2002.
3. T. Inoue, K. Okada, and Y. Matsushita, Learning from TV Programs: Application of TV Presentation to a Videoconferencing System, *Proc. of ACM UIST '95*, pp.147-154, 1995.
4. M. Glenny, R. Tayler (eds), S. M. Eisenstein Selected Works Volume 2, Towards a Theory of Montage, *British Film Institute*, 1991.
5. D. Arijion, Grammar of the Film Language, *Silman-James Press*, Los Angeles, 1976.
6. B. Reeves, C. Nass, The Media Equation, *CSLI Publication*, 1996.
7. A. Kendon, Some Function of Gaze-direction in Social Interaction, *Act. Psychologica*, 26, pp.22-63, 1967.
8. T. Ohno, N. Mukawa, and S. Kawato, Just Blink Your Eyes: A Head-Free Gaze Tracking System, *Ext. abstracts of CHI '03*, pp.950-951, 2003.
9. R. V. Exline, Exploration in The Process of Person Perception: Visual Interaction in Relation to Competition, Sex and Need for Affiliation, *J. of Personality*, 31, pp.1-20, 1963.